erc
European Research Council
Established by the European Commission

Universität
Konstanz

Changes in Null Subjects in Latin American Spanish: A Diachronic Corpus Study

Gemma McCarley

CODILI 2022, 06.10.22

STARFISH
SOCIOLINGUISTIC TYPOLOGY
AND RESPONSIVE FEATURES
IN SYNTACTIC HISTORY

Background

- **Spanish is a null subject language (NSL) which means it can have sentences like (1) that are perfectly grammatical**

1. Spanish [consistent NSL]: (Nosotros) queremos ir a la playa
   English [non-NSL (NNSL)]: *(We) want to go to the beach

- **It's been noticed that in Latin American Spanish (LAS) overt pronouns are being used at higher rates (e.g. Dominican Spanish: Toribio 2000)**
- **This could potentially represent an incipient process towards becoming a NNSL (Camacho 2013)**
- **In the literature, nullness has historically been linked with inversion, e.g. the NSP, because most consistent NSLs like Italian and Spanish also allow inversion (Rizzi 1982, 1986)**
- **This theoretical correlation tracks with findings that SV word order is also on the rise in varieties where overtness is too (Toribio 2000)**

2. Papi, ¿qué ese letrero dice?
   (cf. Papi, ¿qué dice ese letrero?)
   'Daddy, what does that sign say?' (Toribio 2000: 322)

- **Why might this be? One of the biggest characteristics of LAS is its history of significant language contact**

Background: Null Subject Acquisition & Simplification

- When we talk about language contact, we are really talking about language acquisition.
- It has been well-noted in the acquisition literature that null subjects are harder to acquire, particularly for L2 speakers (Bini 1993, Pérez-Leroux & Glass 1999, Margaza & Bel 2006, Sorace 2011, Tsimpli & Lavidas 2019)
- In that case, increasing the use of overt pronouns seems to be an act of simplification
- Language contact, then, is often an impetus for simplification when the simplifying feature is difficult to acquire. Especially when that contact takes the form of short-term, loose-knit, adult language learning (Trudgill 2011, Walkden & Breitbarth 2019)
- That is exactly the context for African learners of Spanish in colonial Latin America

Background: AHLAs

- Specifically, during the colonial period enslaved Africans were brought over to Latin America.
- These adult learners of L2 Spanish might have struggled acquiring the L2-difficult null subject system, preferring overt pronouns (and SV word order).
- Their children would then have nativized this system. This is exactly the scenario Sandro Sessarego (2013) proposes for Latin American Spanish where AHLAs are these nativized varieties.
- So, the next step would be to look into the diachronic trajectory of pronoun realization and word order in Latin American Spanish. I'm in the process of creating a corpus of 60+ texts to do just that.

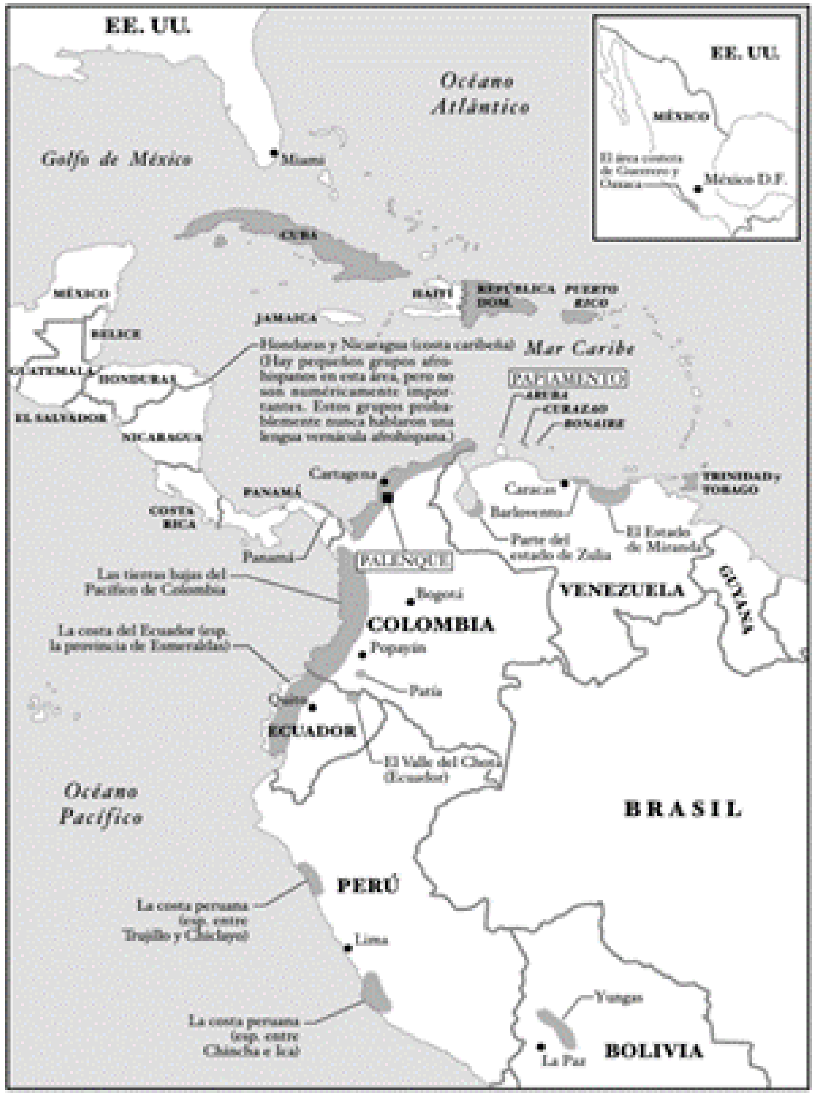

EE. UU.
Océano Atlántico
Golfo de México
Miami
EE. UU.
MÉXICO
El área costera de Guerrero y Oaxaca
México D.F.
CUBA
MÉXICO
HAITÍ
REPÚBLICA DOM.
PUERTO RICO
JAMAICA
BELICE
GUATEMALA
HONDURAS
EL SALVADOR
NICARAGUA
Honduras y Nicaragua (costa caribeña) (Hay pequeños grupos afrohispanos en esta área, pero no son numéricamente importantes. Estos grupos probablemente nunca hablaron una lengua vernácula afrohispana.)
Mar Caribe
PAPIAMENTO
ARUBA
CURAZAO
BONAIRE
Cartagena
PANAMÁ
COSTA RICA
Panamá
PALENQUE
Caracas
Barlovento
Parte del estado de Zulia
El Estado de Miranda
TRINIDAD y TOBAGO
GUYANA
VENEZUELA
Las tierras bajas del Pacífico de Colombia
Bogotá
COLOMBIA
La costa del Ecuador (esp. la provincia de Esmeraldas)
Popayán
Patía
Quito
ECUADOR
El Valle del Chota (Ecuador)
Océano Pacífico
BRASIL
PERÚ
La costa peruana (esp. entre Trujillo y Chiclayo)
Lima
La costa peruana (esp. entre Chincha e Ica)
Yungas
La Paz
BOLIVIA

Figure 1: Afro-Hispanic areas of Latin America (Klee & Lynch 2009:6)

Research Questions

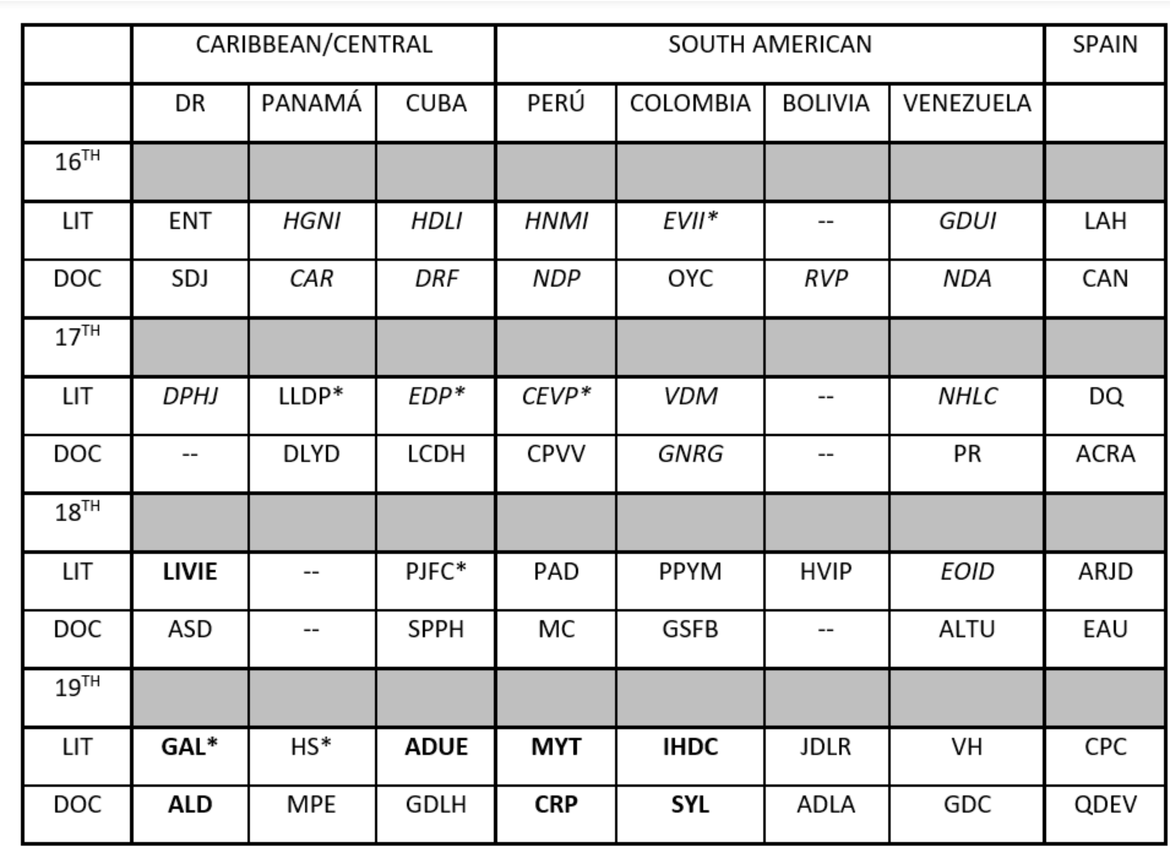
1. *do overtness and SV word order increase diachronically?*

2. *do they have higher rates from Spain > South America > Caribbean?*

3. *do certain genres have higher rates than others?*

Methodology: Corpus

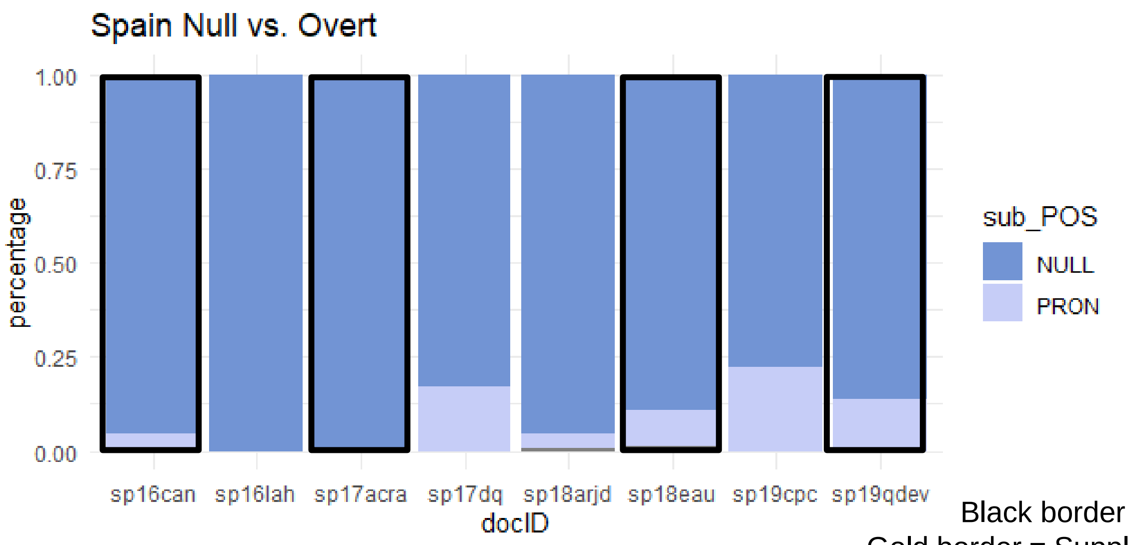
- **This is the main historical corpus covering 57 texts (~2-3k words each) from 8 countries during the 16th-19th centuries**
  - I selected 7 countries from the Caribbean and Central and South America (plus Spain as a control)
  - They were selected for their high Afro-Hispanic populations
- **For each century + country combination, there are ideally 2 texts, one from each genre:**
  - Literature (e.g. novels, plays, poetry)
  - Documents (e.g. newspapers, legal documents, letters)
- **In addition to this corpus, I have also set aside:**
  - A transcript of an interview in Afro-Bolivian from 2010
- **The main sources for the texts are Cervantes Virtual, dLOC, and BDH**
- **Each text has been transcribed by myself or my research assistant, parsed by the Stanford Parser, and then annotated by hand**

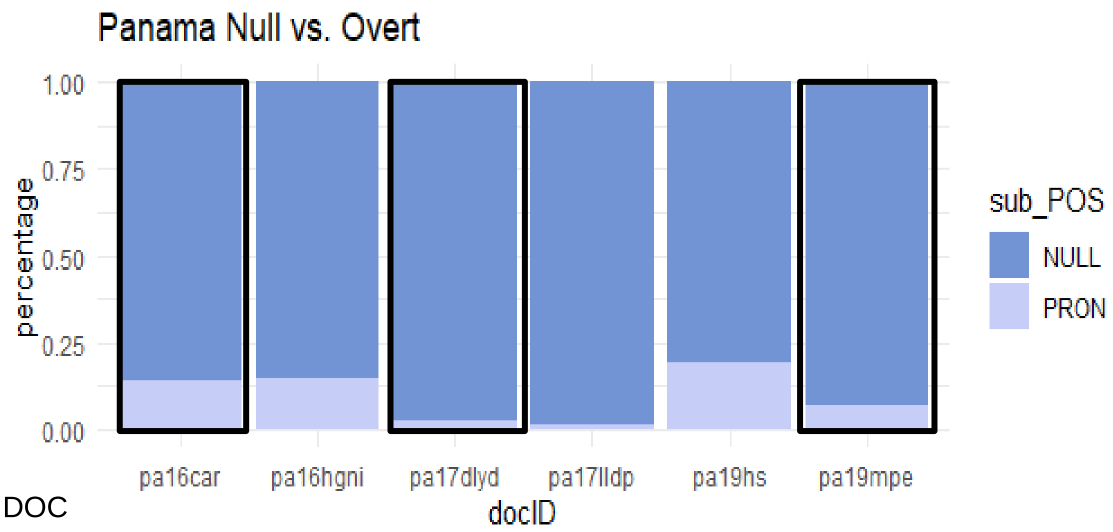


	CARIBBEAN/CENTRAL			SOUTH AMERICAN				SPAIN
	DR	PANAMÁ	CUBA	PERÚ	COLOMBIA	BOLIVIA	VENEZUELA	
16TH								
LIT	ENT	*HGNI*	*HDLI*	*HNMI*	*EVII**	--	*GDUI*	LAH
DOC	SDJ	*CAR*	*DRF*	*NDP*	OYC	*RVP*	*NDA*	CAN
17TH								
LIT	*DPHJ*	LLDP*	*EDP**	*CEVP**	*VDM*	--	*NHLC*	DQ
DOC	--	DLYD	LCDH	CPVV	*GNRG*	--	PR	ACRA
18TH								
LIT	**LIVIE**	--	PJFC*	PAD	PPYM	HVIP	*EOID*	ARJD
DOC	ASD	--	SPPH	MC	GSFB	--	ALTU	EAU
19TH								
LIT	**GAL***	HS*	**ADUE**	**MYT**	**IHDC**	JDLR	VH	CPC
DOC	**ALD**	MPE	GDLH	**CRP**	**SYL**	ADLA	GDC	QDEV

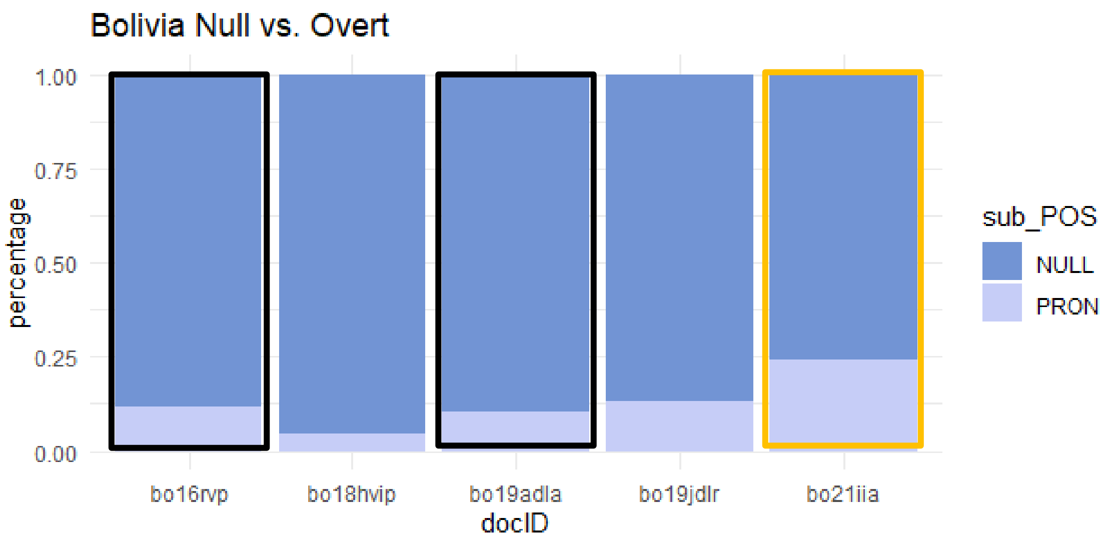
Table 1: Corpus Composition | **AH** | *Born in Spain* | Verse*

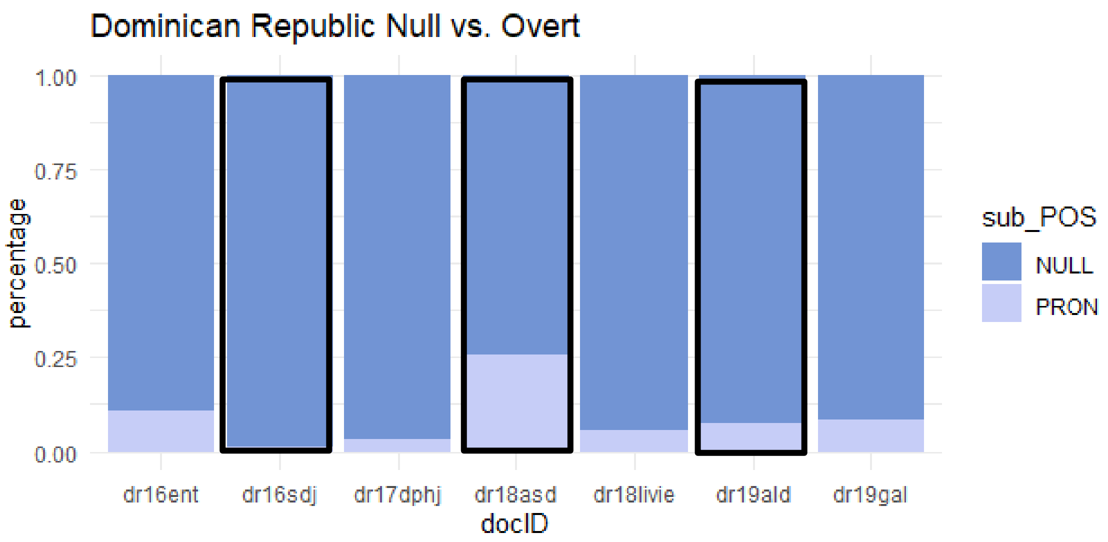
Pronoun Realization (Percent)



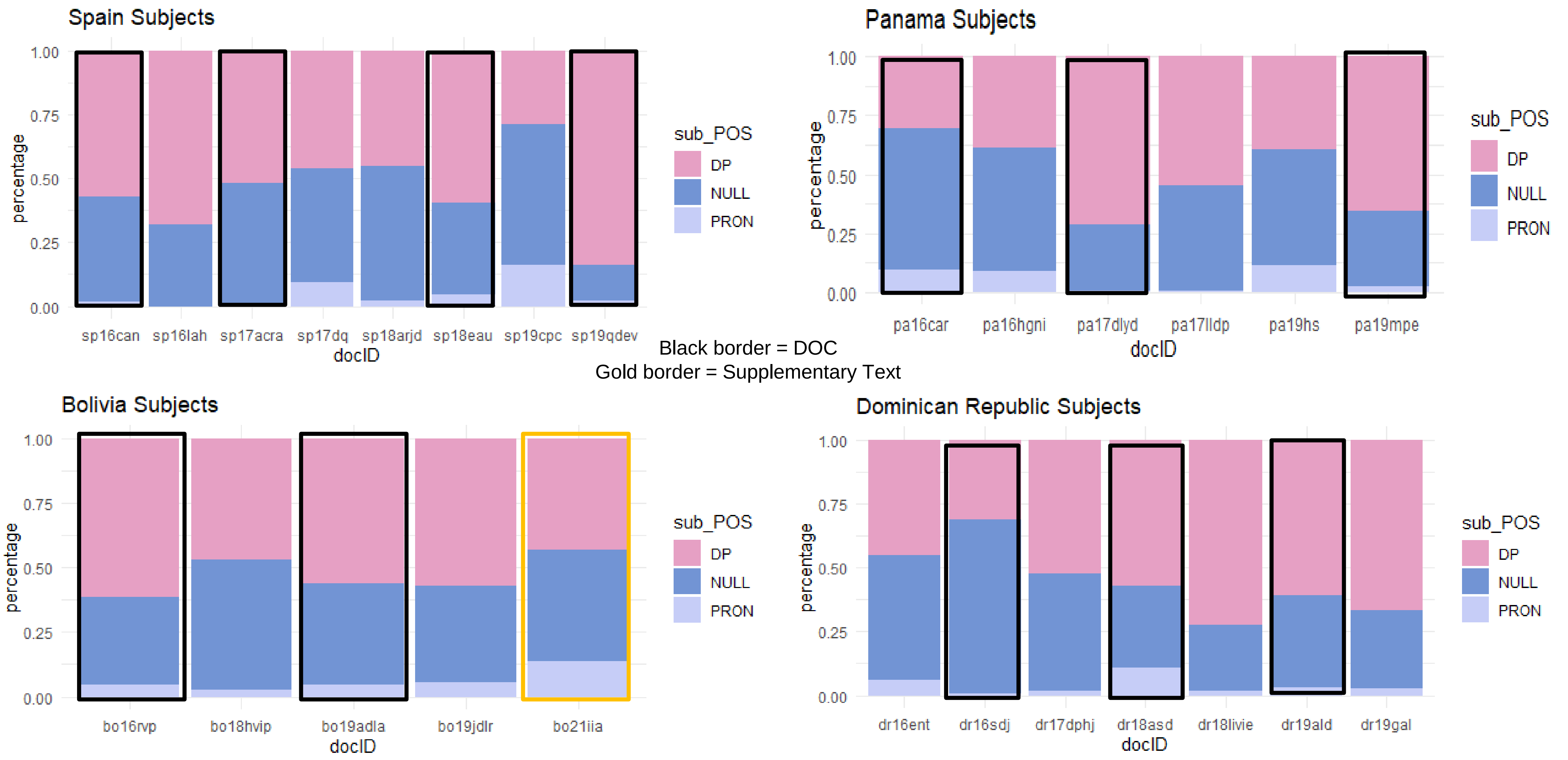


Black border = DOC
Gold border = Supplementary Text



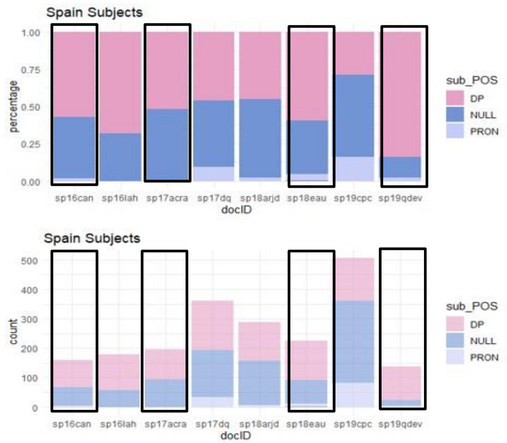


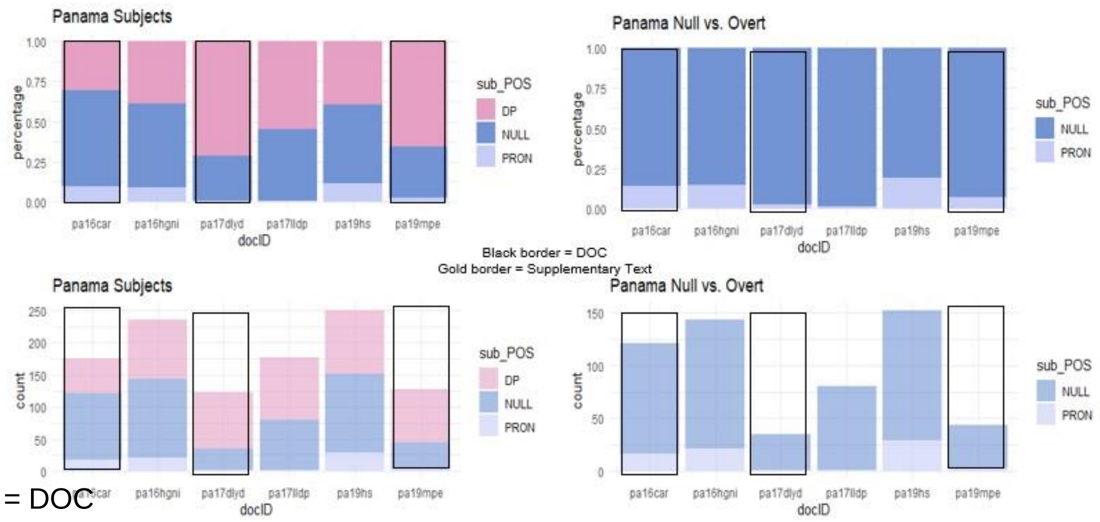
Subject Realization (Percent)



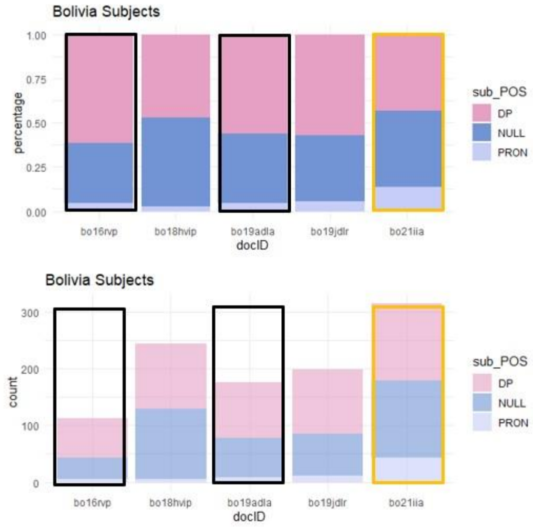
Black border = DOC
Gold border = Supplementary Text

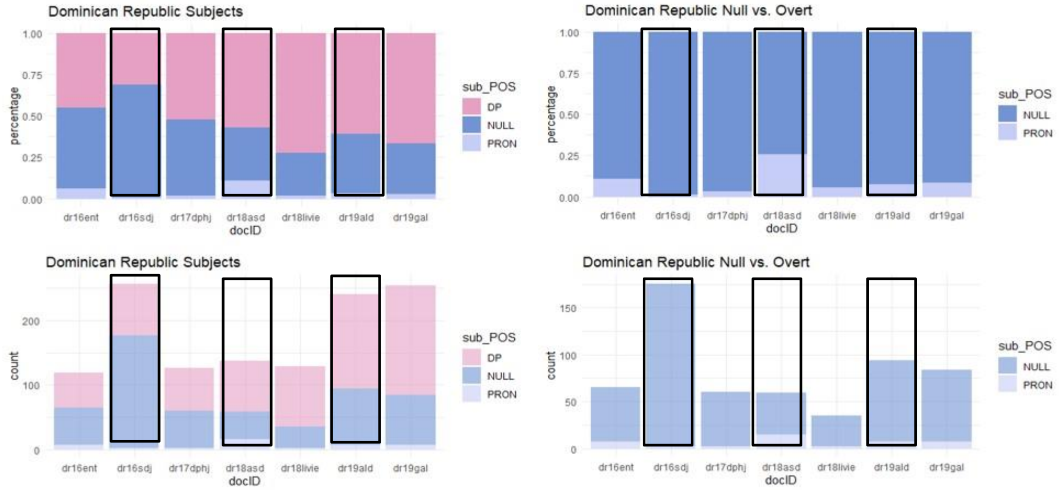
Subject Realization (Count)



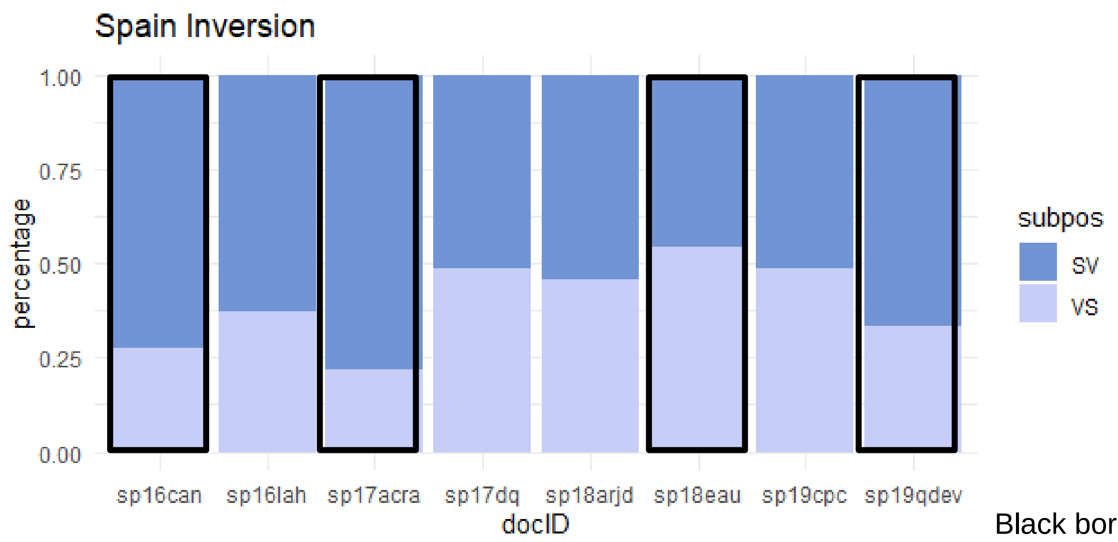


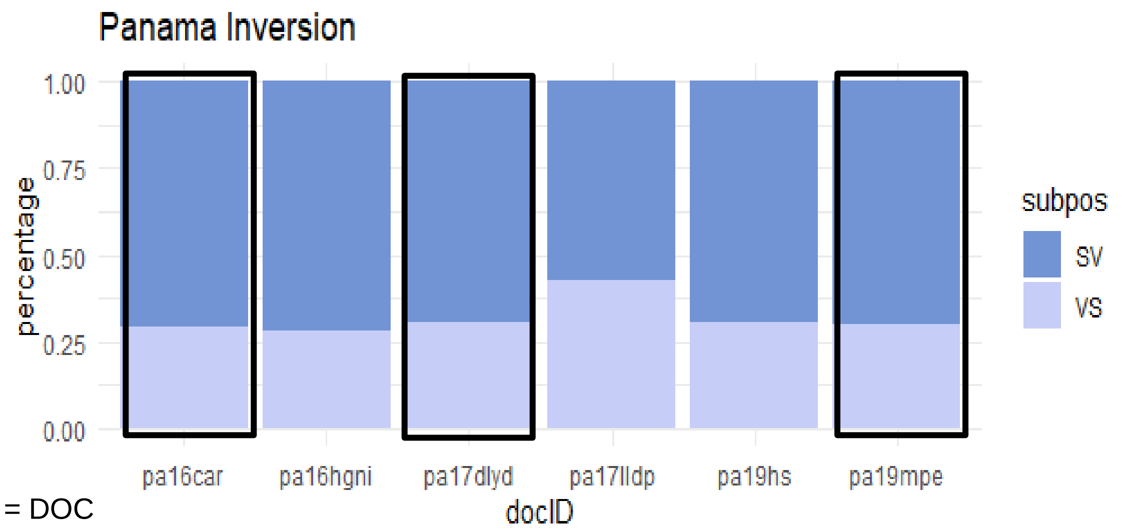
Black border = DOC
Gold border = Supplementary Text



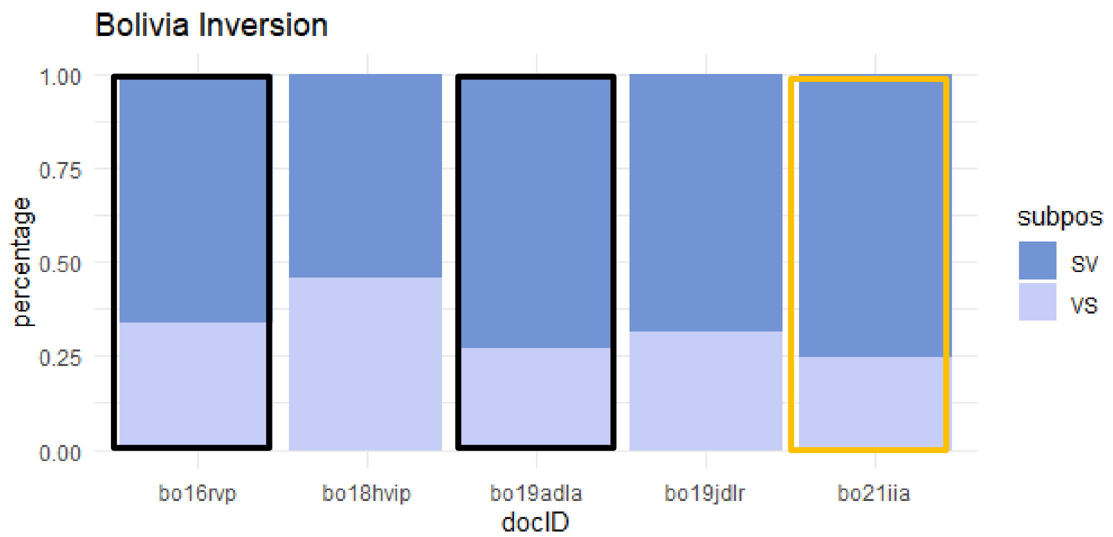


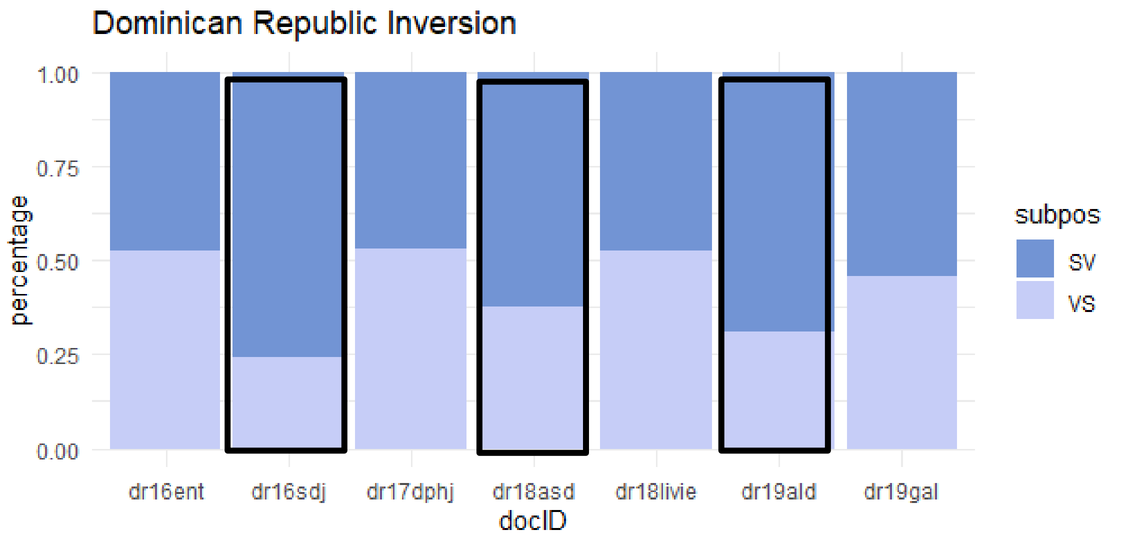
Word Order (Percent)


Spain Inversion

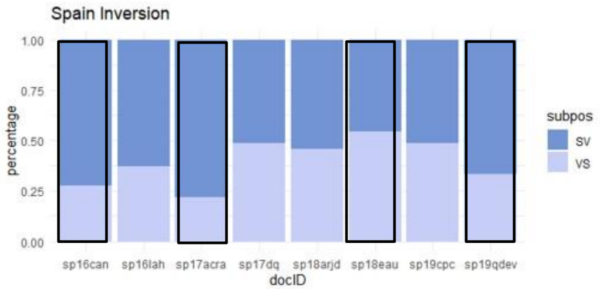

Panama Inversion

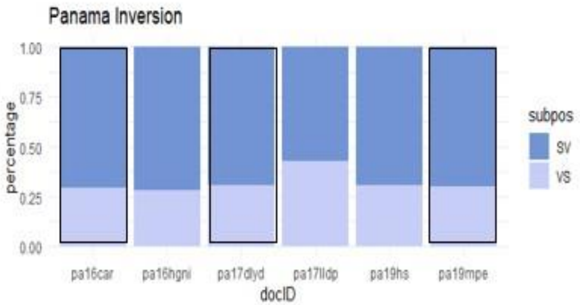
Black border = DOC
Gold border = Supplementary Text

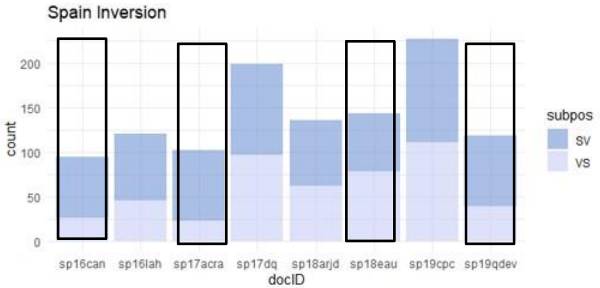

Bolivia Inversion

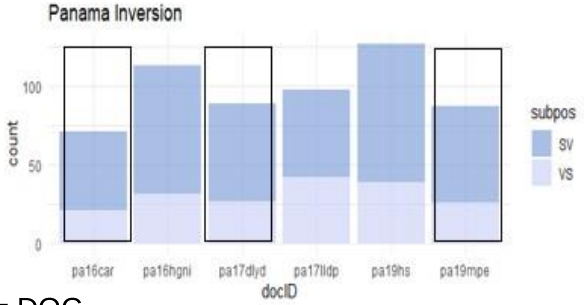

Dominican Republic Inversion

Word Order (count)

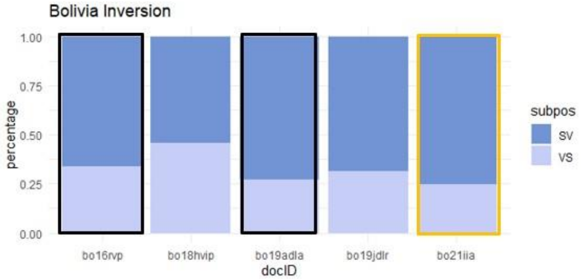


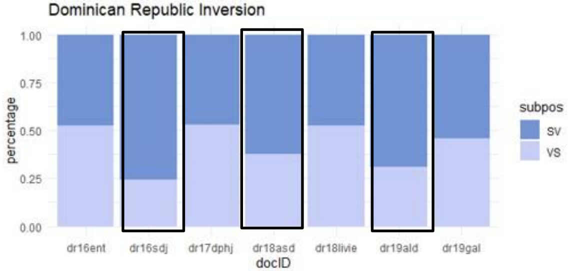


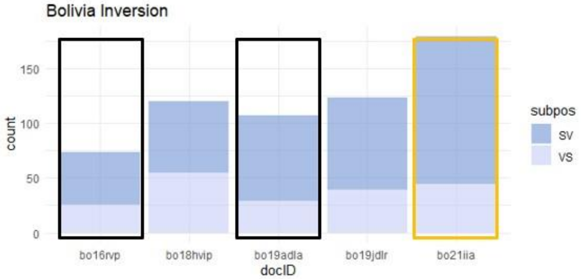


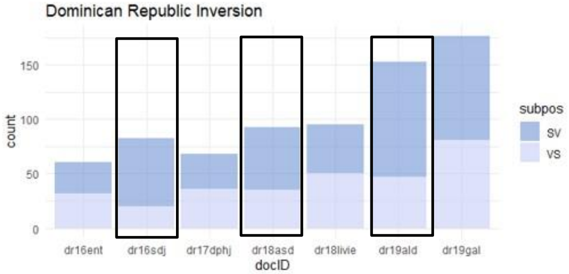


Black border = DOC
Gold border = Supplementary Text

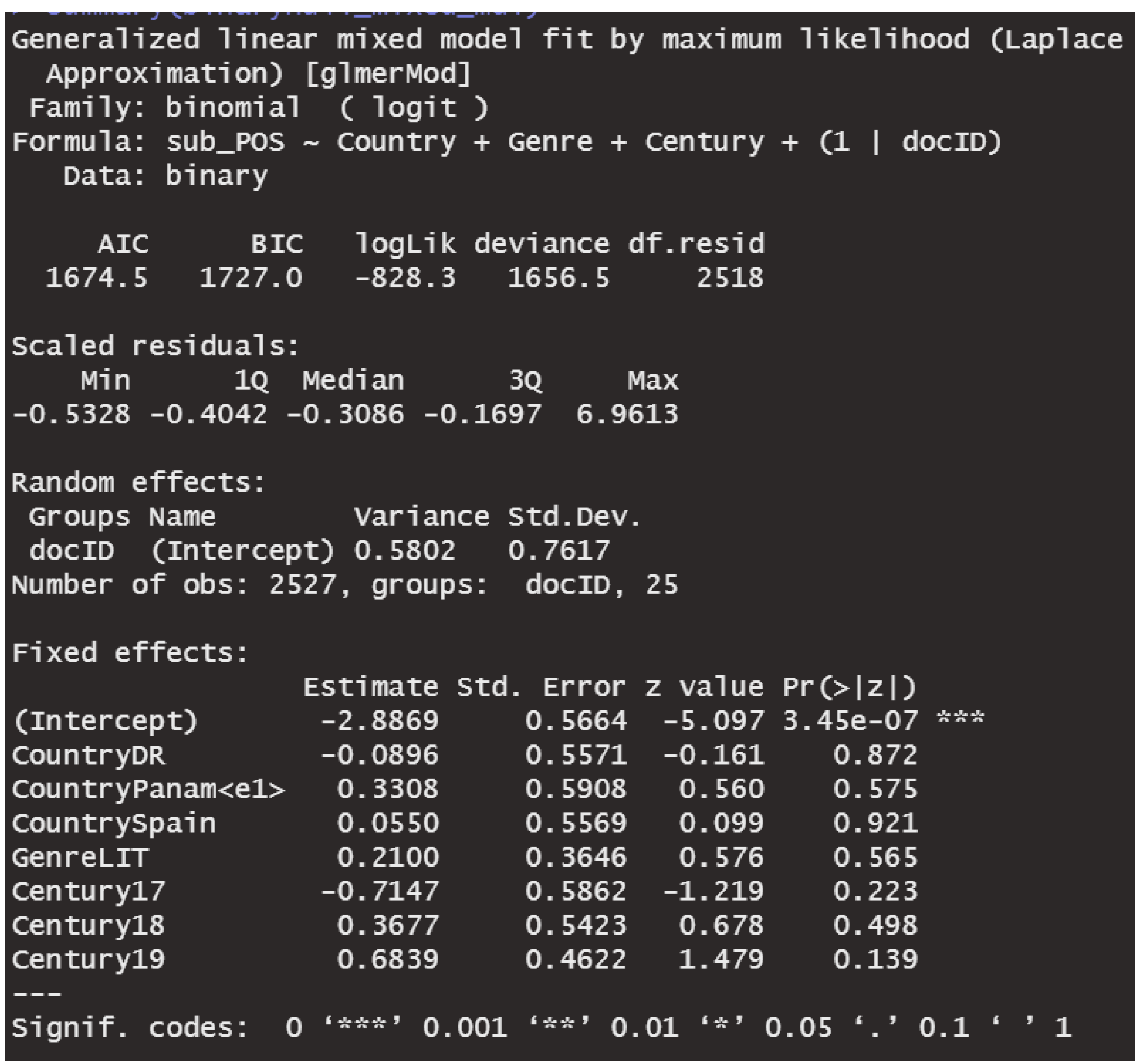








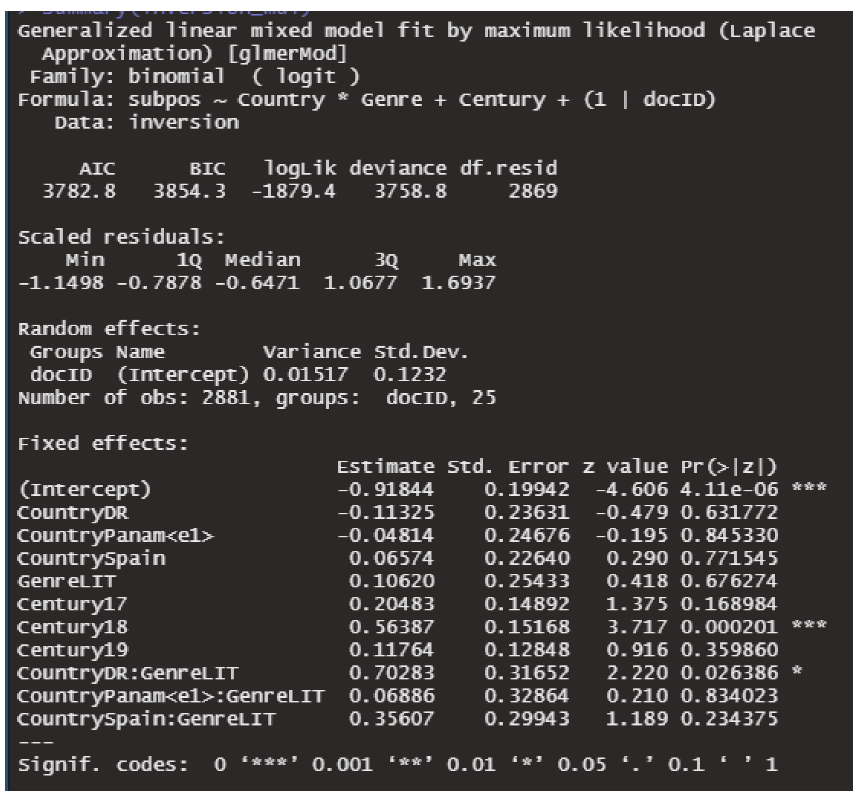
Mixed Models: Pronoun Realization



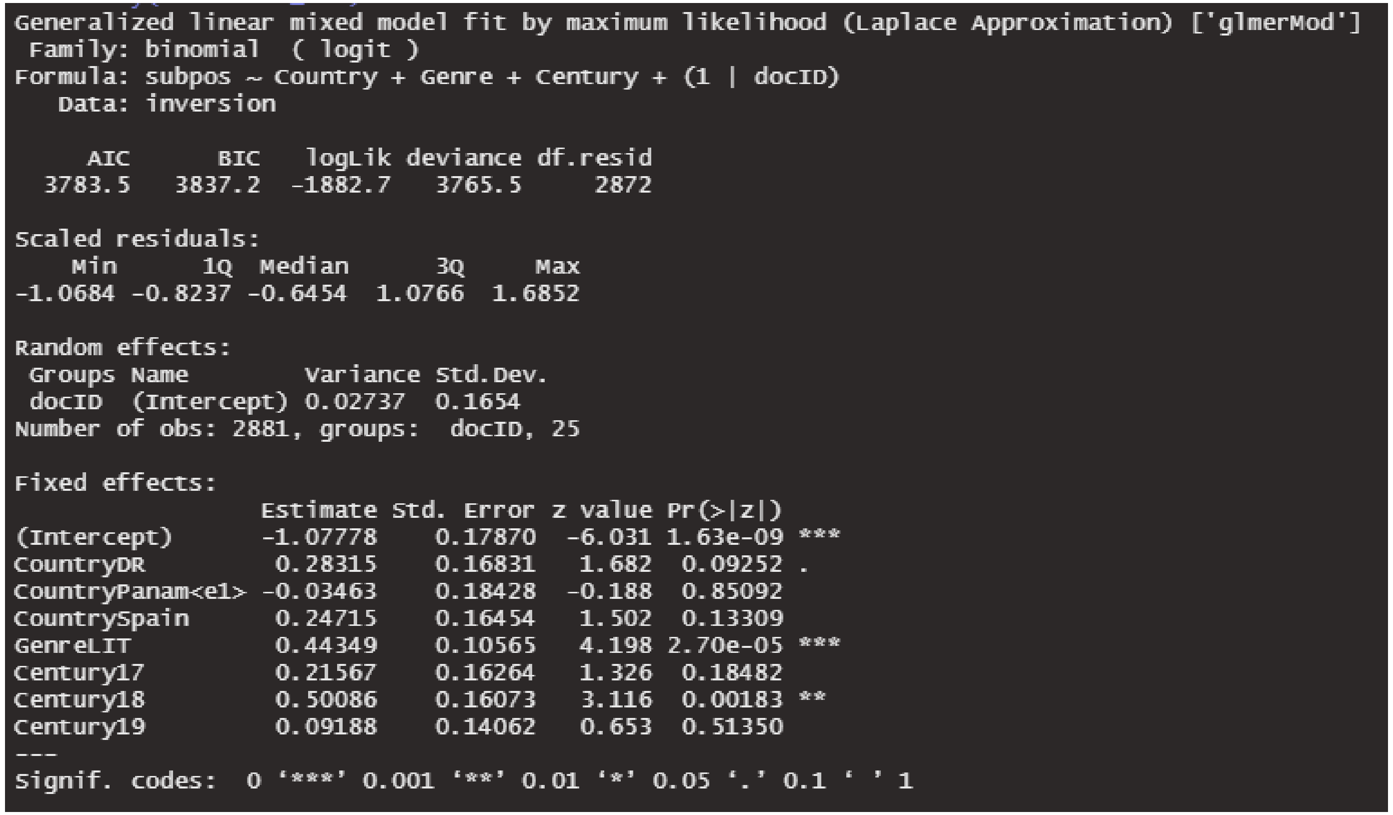
```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: sub_POS ~ Country + Genre + Century + (1 | docID)
   Data: binary

     AIC      BIC   logLik deviance df.resid
  1674.5   1727.0   -828.3   1656.5     2518

Scaled residuals:
    Min      1Q  Median      3Q     Max
-0.5328 -0.4042 -0.3086 -0.1697  6.9613

Random effects:
 Groups Name        Variance Std.Dev.
 docID  (Intercept) 0.5802   0.7617
Number of obs: 2527, groups:  docID, 25

Fixed effects:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.8869     0.5664  -5.097 3.45e-07 ***
CountryDR         -0.0896     0.5571  -0.161    0.872
CountryPanam<e1>   0.3308     0.5908   0.560    0.575
CountrySpain       0.0550     0.5569   0.099    0.921
GenreLIT           0.2100     0.3646   0.576    0.565
Century17         -0.7147     0.5862  -1.219    0.223
Century18          0.3677     0.5423   0.678    0.498
Century19          0.6839     0.4622   1.479    0.139
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- **Models**
  - glmer from lme4 package in R
  - Looking at the fixed variables of Country, Genre, and Century and their interactions for pronoun realization and word order
  - Neither model would converge with Year as continuous variable (even when used as the only variable)
- **Pronoun Realization**
  - Country*Genre*Century : **no**
  - Country*Genre + Century : **yes (nothing close to significant)**
  - Country + Genre + Century : **yes (nothing significant)**
  - Country / Genre / Century: **yes (still nothing significant)**
  - So, the model doesn’t find anything.
  - We’ll see if that changes once the corpus is complete and there’s more data

Mixed Models: Word Order



```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: subpos ~ Country * Genre + Century + (1 | docID)
   Data: inversion

     AIC      BIC   logLik deviance df.resid
  3782.8   3854.3  -1879.4   3758.8     2869

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.1498 -0.7878 -0.6471  1.0677  1.6937

Random effects:
 Groups Name        Variance Std.Dev.
 docID  (Intercept) 0.01517  0.1232
Number of obs: 2881, groups:  docID, 25

Fixed effects:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -0.91844    0.19942  -4.606 4.11e-06 ***
CountryDR                 -0.11325    0.23631  -0.479 0.631772
CountryPanam<e1>          -0.04814    0.24676  -0.195 0.845330
CountrySpain               0.06574    0.22640   0.290 0.771545
GenreLIT                   0.10620    0.25433   0.418 0.676274
Century17                  0.20483    0.14892   1.375 0.168984
Century18                  0.56387    0.15168   3.717 0.000201 ***
Century19                  0.11764    0.12848   0.916 0.359860
CountryDR:GenreLIT         0.70283    0.31652   2.220 0.026386 *
CountryPanam<e1>:GenreLIT  0.06886    0.32864   0.210 0.834023
CountrySpain:GenreLIT      0.35607    0.29943   1.189 0.234375
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

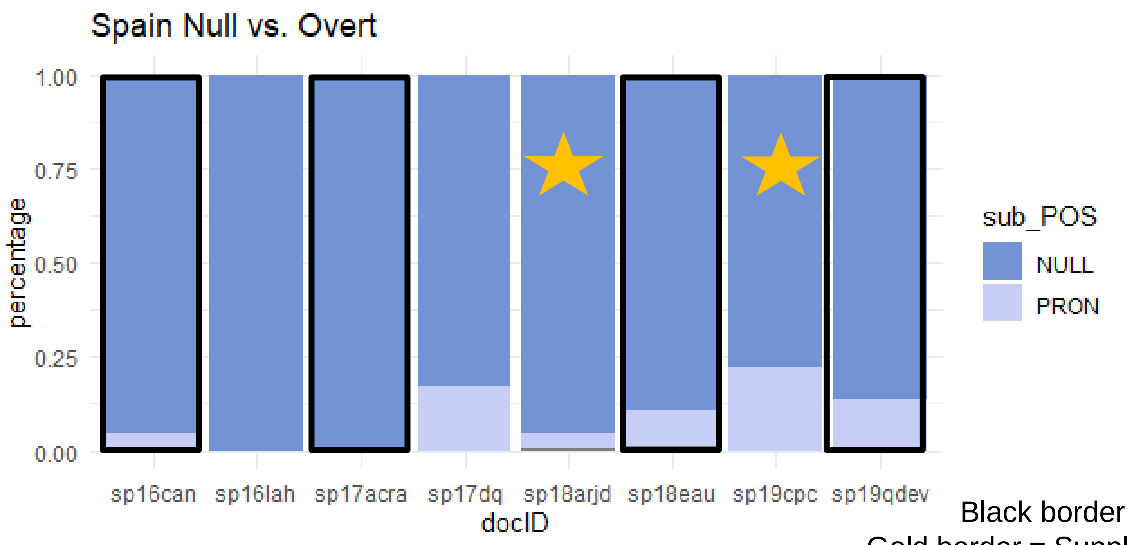


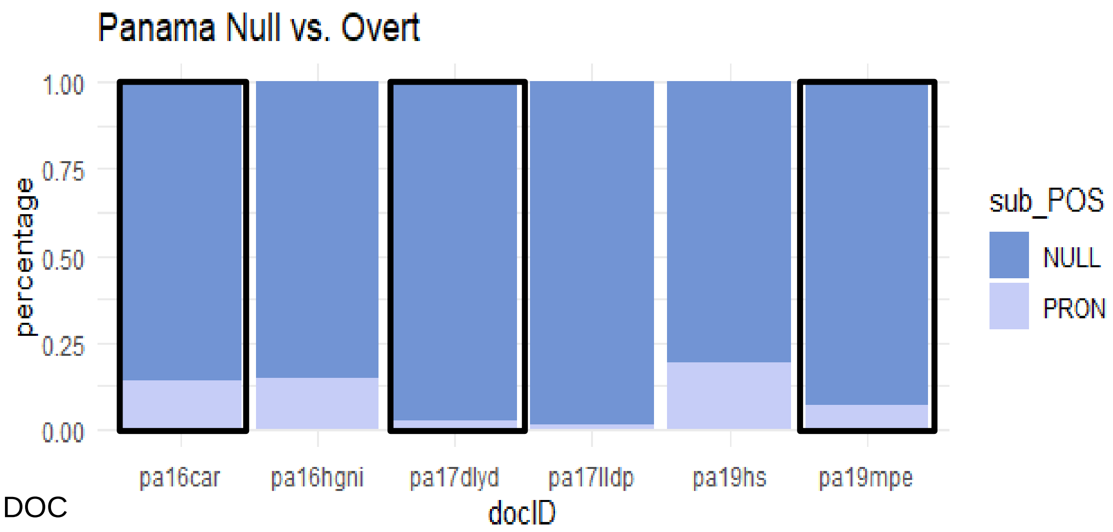
```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: binomial  ( logit )
Formula: subpos ~ Country + Genre + Century + (1 | docID)
   Data: inversion

     AIC      BIC   logLik deviance df.resid
  3783.5   3837.2  -1882.7   3765.5     2872

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.0684 -0.8237 -0.6454  1.0766  1.6852

Random effects:
 Groups Name        Variance Std.Dev.
 docID  (Intercept) 0.02737  0.1654
Number of obs: 2881, groups:  docID, 25

Fixed effects:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.07778    0.17870  -6.031 1.63e-09 ***
CountryDR         0.28315    0.16831   1.682  0.09252 .
CountryPanam<e1> -0.03463    0.18428  -0.188  0.85092
CountrySpain      0.24715    0.16454   1.502  0.13309
GenreLIT          0.44349    0.10565   4.198 2.70e-05 ***
Century17         0.21567    0.16264   1.326  0.18482
Century18         0.50086    0.16073   3.116  0.00183 **
Century19         0.09188    0.14062   0.653  0.51350
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Country*Genre*Century : **no**
- Country*Genre + Century : **yes**
  - **18th century**
  - **interaction between Genre and DR**

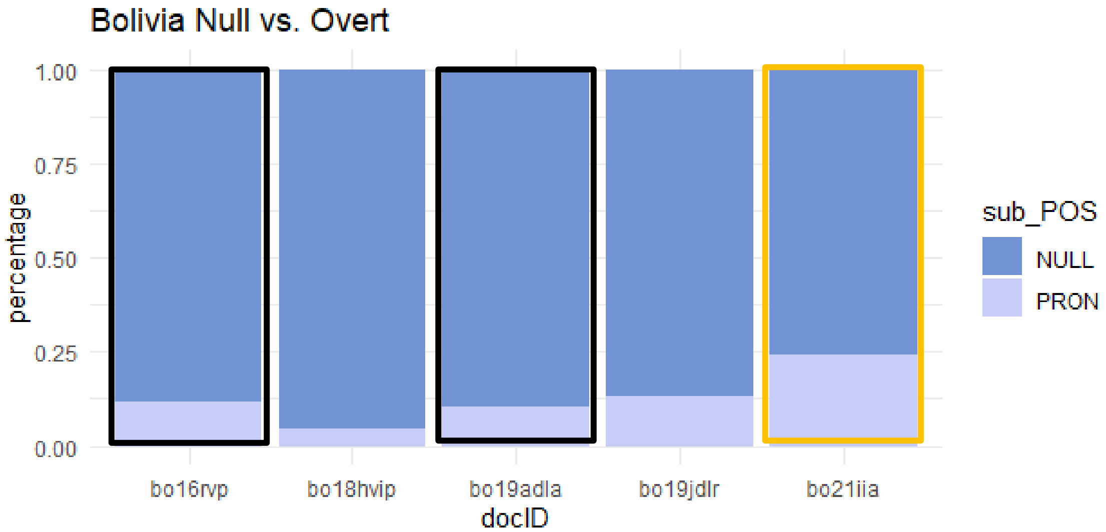
- Country + Genre + Century **: yes**
  - **18th still but less so**
  - **Genre in general**
  - Same results when each variable run individually
- Why the 18th century? I can't say other than that since year had to be adjusted to century, the model doesn't take into account that there's a diachronic relationship

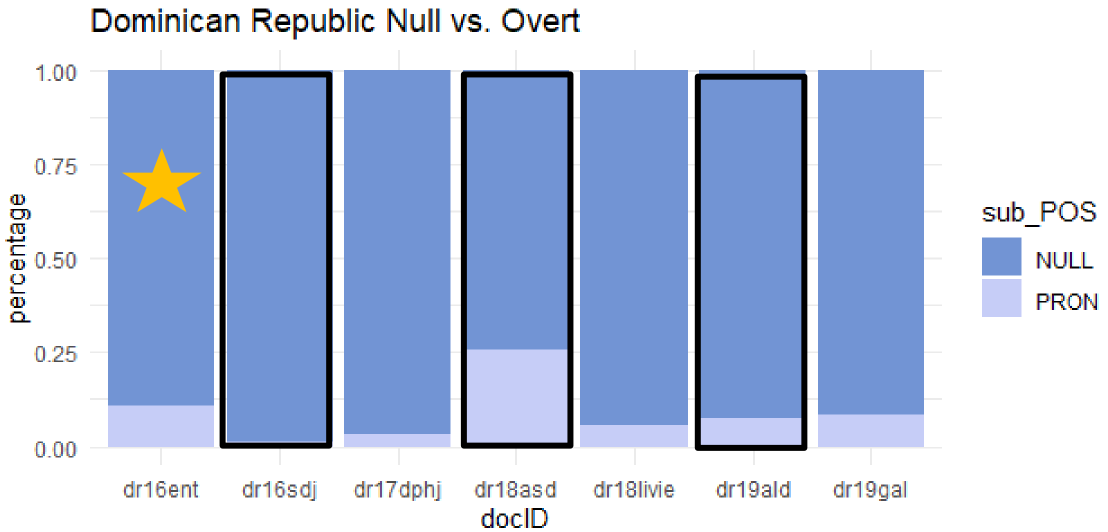
Pronoun Realization (Percent) -- PLAYS


Spain Null vs. Overt


Panama Null vs. Overt

Black border = DOC
Gold border = Supplementary Text


Bolivia Null vs. Overt


Dominican Republic Null vs. Overt

Measuring Orality

•Rosemeyer (2019) measured orality levels in a diachronic corpus of Brazilian Portuguese plays:

- The plays followed a shift toward reflecting spoken speech over the centuries

•Rosemeyer (2019) variables:

•Present progressive

•Demonstrative neuter pronouns

•Time and place adverbs

•Discourse markers

•Private verbs

•My variables:

•Progressive

•Demonstrative neuter pronouns

- *esto/eso/aquello*

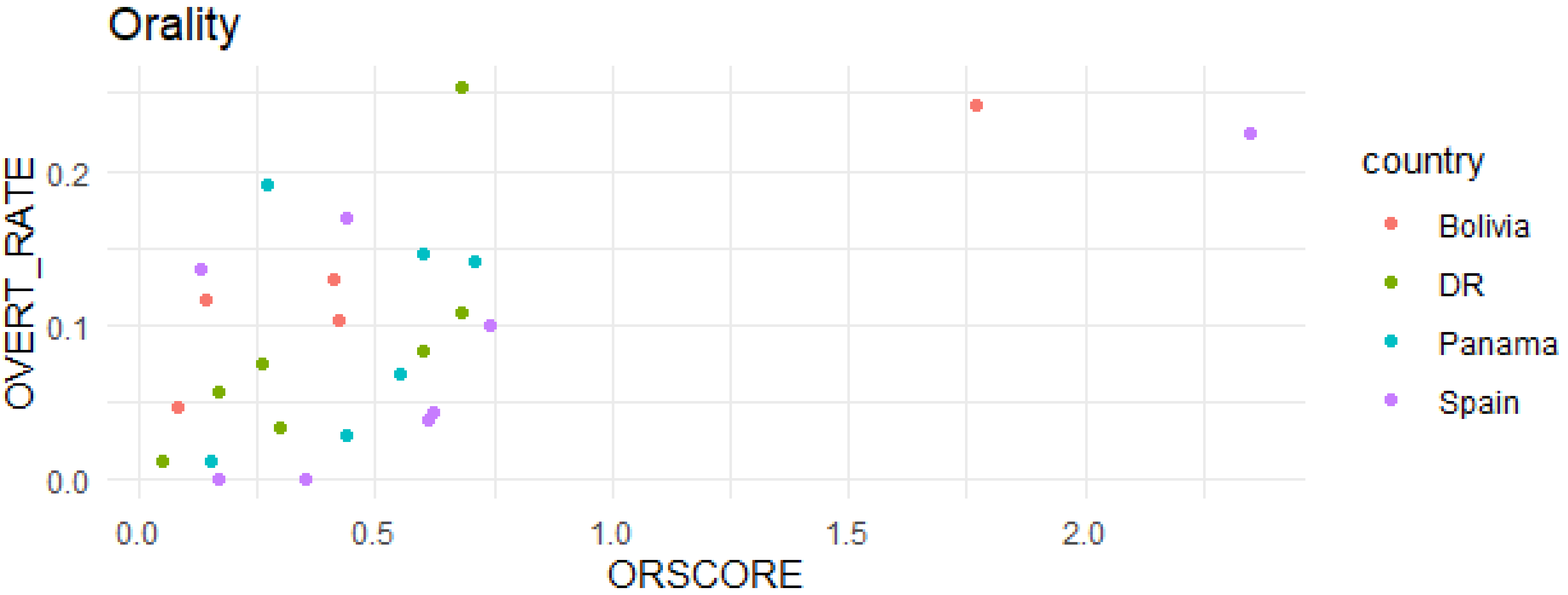
•Time and place adverbs

- *aqui/ahora*

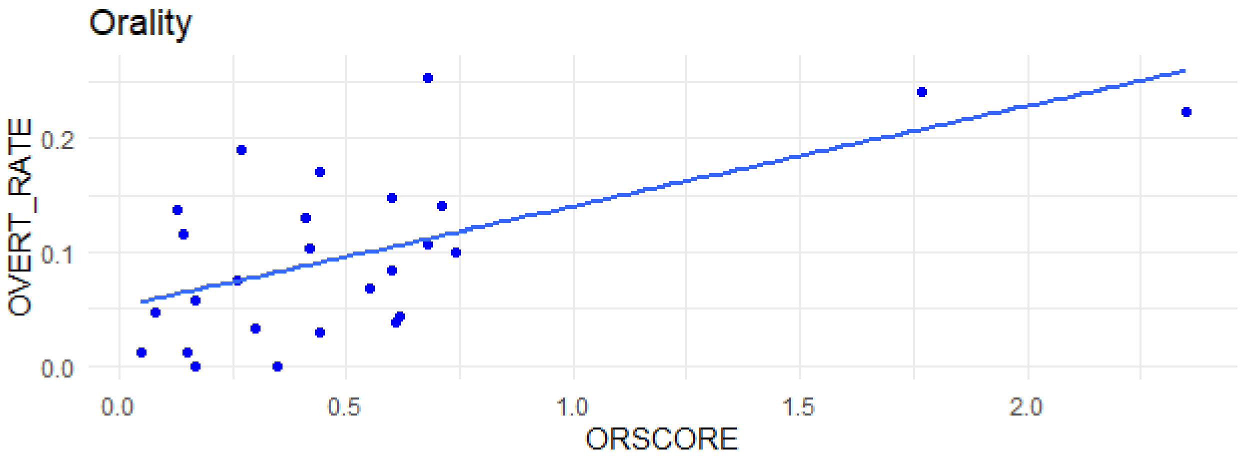
•Private verbs

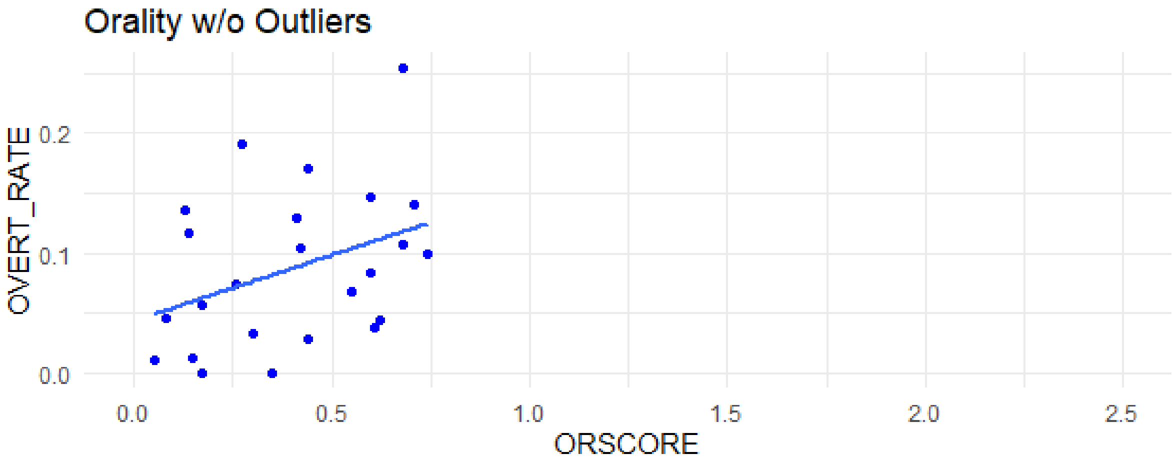
- *pensar* ‘to think’ / *creer* ‘to believe'

Plotting Orality Against Overtness Rates

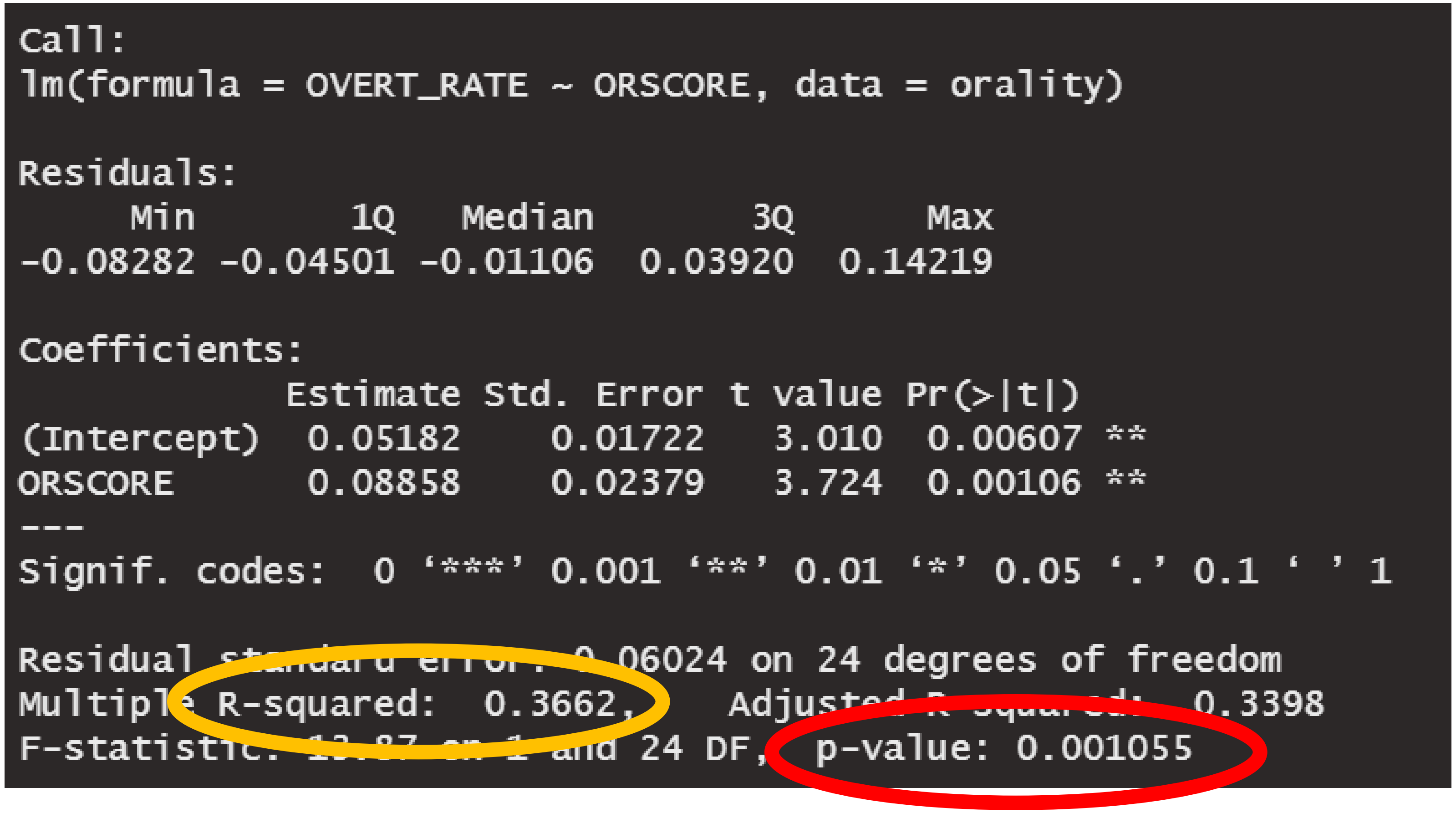


Plotting Orality Against Overtness Rates: Regression



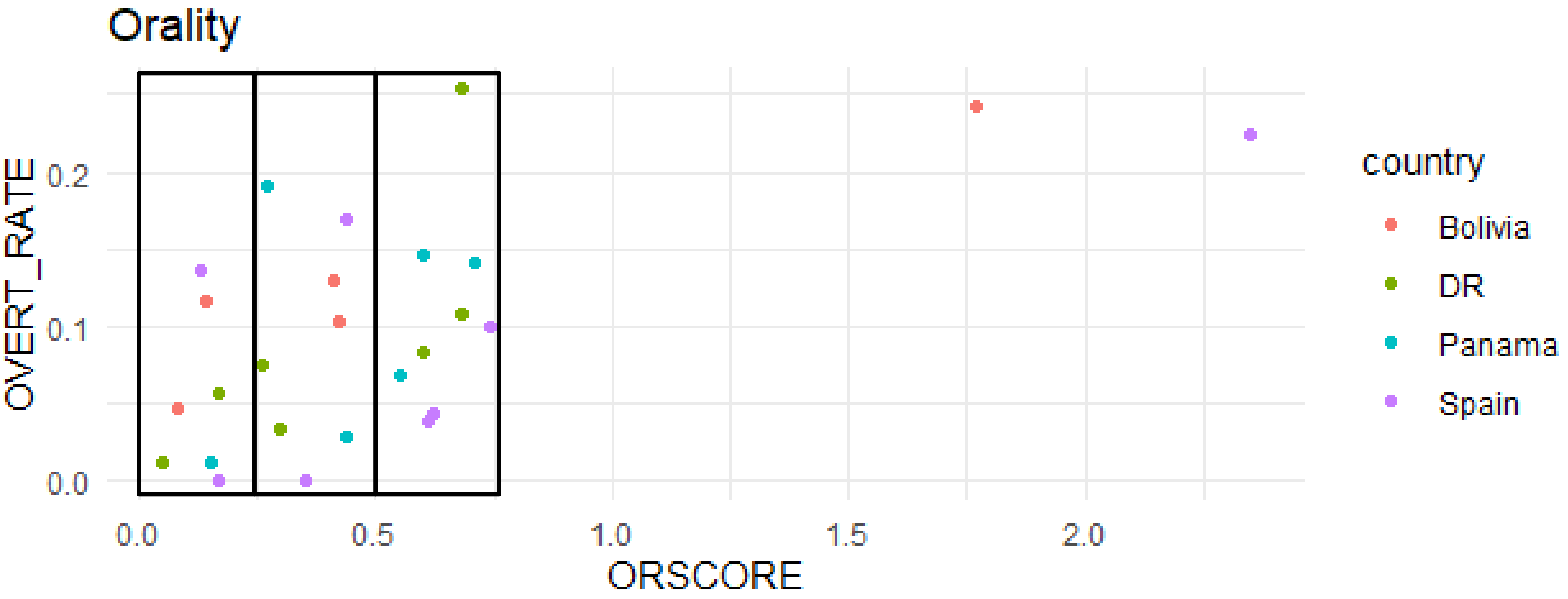


Modelling Orality

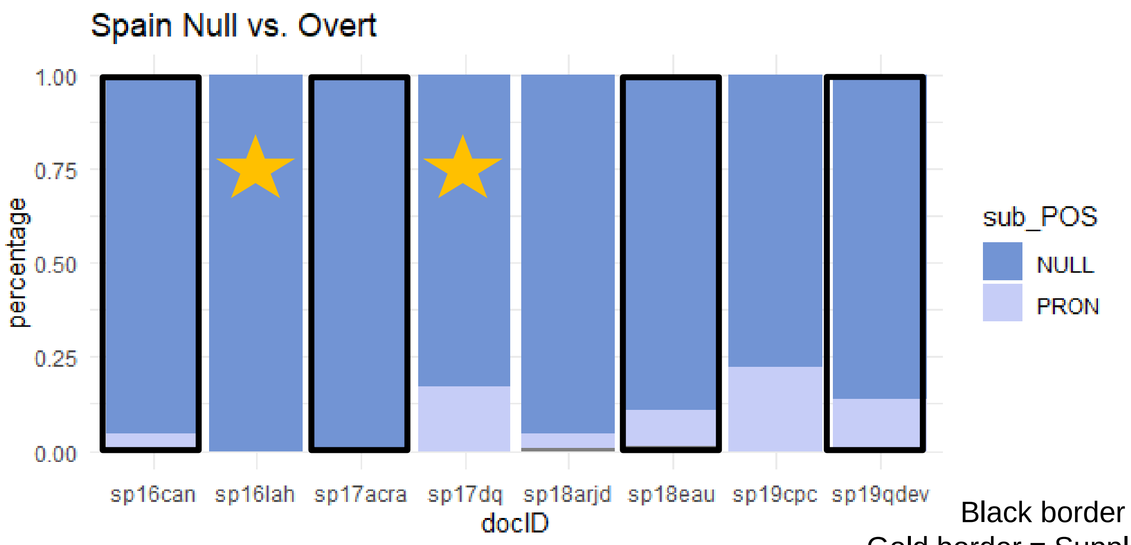


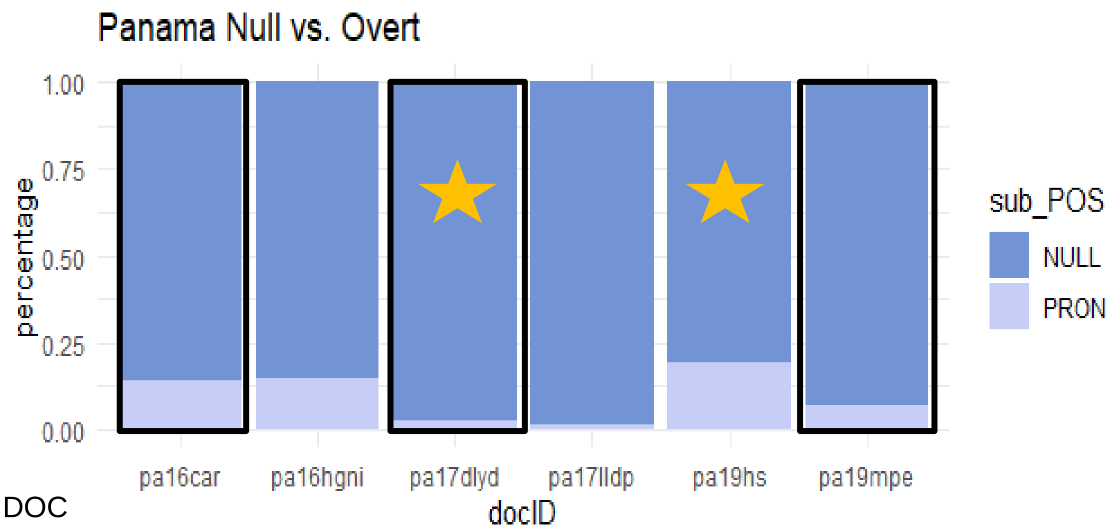
```
Call:
lm(formula = OVERT_RATE ~ ORSCORE, data = orality)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.08282 -0.04501 -0.01106  0.03920  0.14219 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  0.05182    0.01722   3.010  0.00607 **
ORSCORE      0.08858    0.02379   3.724  0.00106 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06024 on 24 degrees of freedom
Multiple R-squared:  0.3662,    Adjusted R-squared:  0.3398 
F-statistic: 13.87 on 1 and 24 DF,  p-value: 0.001055
```

Plotting Orality Against Overtness Rates

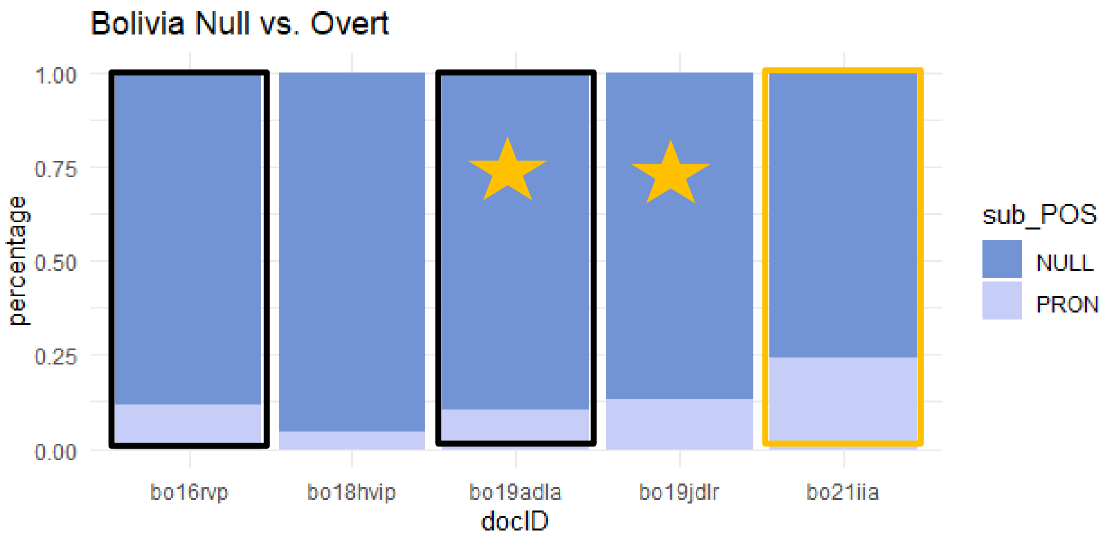


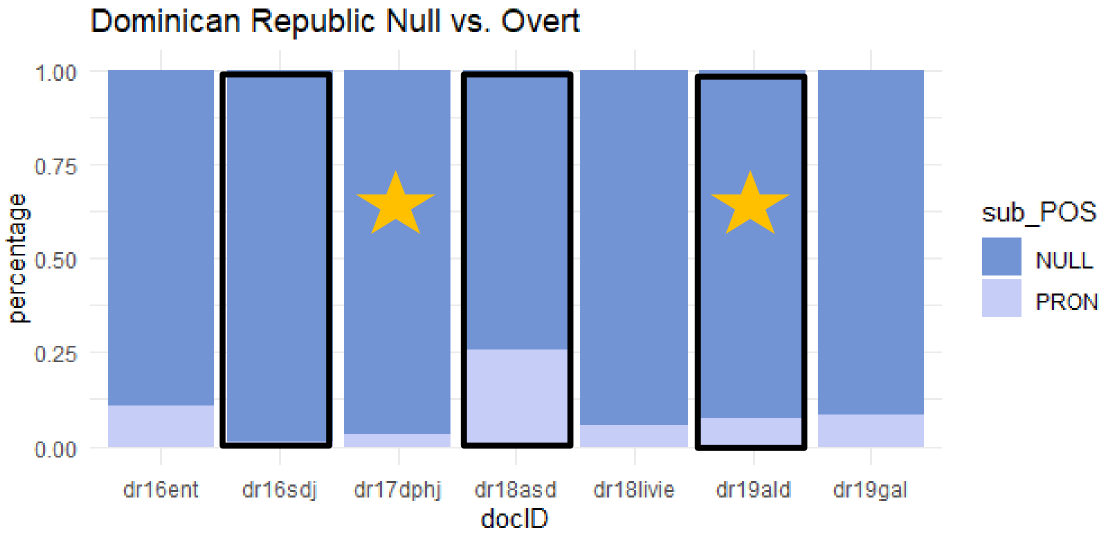
Pronoun Realization (Percent) – Mid-Orality





Black border = DOC
Gold border = Supplementary Text





Putting the Models into Perspective

	CARIBBEAN/CENTRAL			SOUTH AMERICAN				SPAIN
	DR	PANAMÁ	CUBA	PERÚ	COLOMBIA	BOLIVIA	VENEZUELA	
16TH								
LIT	ENT	*HGNI*	*HDLI*	*HNMI*	*EVII**	--	*GDUI*	LAH
DOC	SDJ	*CAR*	*DRF*	*NDP*	OYC	*RVP*	*NDA*	CAN
17TH								
LIT	*DPHJ*	LLDP*	*EDP**	*CEVP**	*VDM*	--	*NHLC*	DQ
DOC	--	DLYD	LCDH	CPVV	*GNRG*	--	PR	ACRA
18TH								
LIT	**LIVIE**	--	PJFC*	PAD	PPYM	HVIP	*EOID*	ARJD
DOC	ASD	--	SPPH	MC	GSFB	--	ALTU	EAU
19TH								
LIT	**GAL***	HS*	**ADUE**	**MYT**	**IHDC**	JDLR	VH	CPC
DOC	**ALD**	MPE	GDLH	**CRP**	**SYL**	ADLA	GDC	QDEV

- **There will be more than double the data by the time the corpus is complete**
- **It is important to keep in mind that this is just preliminary data**
- **When the models have more to work with, they may yield some significant findings**

Conclusion

Main Research Questions:

1. ***do overtness and SV word order increase diachronically?***
2. ***do they have higher rates from Spain > South America > Caribbean?***
3. ***do certain genres have higher rates than others?***

- **Inversion does show a genre effect, preferring "DOC" to "LIT"**
- **Pronominal data had too much inter- and intra- country variation. Why?**
  - Genre? Subgenre?
  - Orality!
- **Significant relationship between overtness and orality**
- **Next steps: figure out a way to account for orality in the corpus in order to further investigate diachronic and regional changes**
- **Hopefully through the mixed model once the rest of the data is ready**

References

Bini, M. (1993). “La adquisición del italiano: Más allá de las propiedades sintácticas delparámetro pro-drop.” In J. M. Liceras (Ed.), *La lingüística y el análisis de los sistemasno nativos:* 126–139. Dovehouse Editions Canada.

Camacho, José. 2013. *Null subjects*. Cambridge: Cambridge University Press.

Cerrón-Palomino, Álvaro. 2018. “Variable subject pronoun expression in Andean Spanish: a drift from the acrolect”. *Onomázein* 1 (42): 53-73.

Klee, C.A. & Lynch, A. 2009. *El español en contacto con otras lenguas*. Washington DC: Georgetown University Press.

Margaza, P., & Bel, A. (2006). “Null subjects at the syntax–pragmatics interface: Evidence from Spanish interlanguage of Greek speakers.” In M. Grantham O’Brien, C. Shea, &J. Archibald (Eds.), *Proceedings of the 8th Generative Approaches to Second Language Acquisition Conference (GASLA 2006):* 88–97. Cascadilla Proceedings Project.

Pérez-Leroux, A. T., & Glass, W. R. (1999). “Null anaphora in Spanish second language acquisition: Probabilistic versus generative approaches.” *Second Language Research*, 15 (2): 220–249.

Rizzi, Luigi. 1982*. Issues in Italian syntax*. Dordrecht: Foris.

Rizzi, Luigi. 1986. Null objects in Italian and the theory of pro. *Linguistic Inquiry* 17: 501–57.

Rosemeyer, Malte. 2019. "Actual and apparent change in Brazilian Portuguese wh-interrogatives.” *Language Variation and Change* 31(2): 165–191. CUP.

Sánchez, M.E. 2008. “Tipos de cláusula, clases verbales y posición del sujeto en español.” *Lexis* XXXII/1, 83-105.

Sessarego, Sandro. 2013. “Afro-Hispanic Contact Varieties as Conventionalized Advanced Second Languages”. *IBERIA* 5 (1): 99-125.

Sorace, Antonella. 2011. “Pinning down the concept of “interface” in bilingualism”. *Linguistic Approaches to Bilingualism* 1(1): 1-33.

Toribio, Almeida J. 2000. “Setting parametric limits on dialectal variation in Spanish”. *Lingua: International Review of General Linguistics* 110 (5): 315–341.

Trudgill, Peter. 2011. Sociolinguistic typology: *Social determinants of linguistic complexity.* Oxford: OUP.

Tsimpli, Iantha Maria and Lavidas, Nikolaos. 2019. “Object Omission in Contact: Object Clitics and Definite Articles in the West Thracian Greek (Evros) Dialect”. *Journal of Language Contact* 12: 141-190.

Walkden, George and Breitbarth, Anne. 2019. "Interpreting (un)interpretability" *Theoretical Linguistics* 45 (3-4): 309-317.

Universität
Konstanz

Thank you
for listening!

STARFISH

SOCIOLINGUISTIC TYPOLOGY
AND RESPONSIVE FEATURES
IN SYNTACTIC HISTORY

gemma-hunter.mccarley@uni-konstanz.de
https://www.ling.uni-konstanz.de/en/walkden/starfish/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 851423

erc
European Research Council
Established by the European Commission

Appendix: Full Tagset

Sentence:

- poem title/letter number (if applicable)
- speaker number/ character name (if applicable)

Subject:

- dep(endency) type: "nsubj" (Nominal Subject)
- subpos (subject position): **SV/VS**
- POS
  - 3p inanimate expletives: PRON-EXP or NULL-EXP
  - relative pronouns: PRON-REL (these get excluded)
  - passive 'se': XPOS-PASS, NULL-PASS*
  - passive 'se' expletives: NULL-EXP-PASS
  - impersonal 'se': NULL-IMP
  - impersonal expletives (e.g. *hay* 'there is/are'): change to NULL-EXP-IMP

Subject pronouns:

- morphology
  - person: 1/2/3/u (u is for 'usted/es')
  - number: s/p/v (v is for 'vos')
  - e.g. "nosotros" = 1p
- psub (previous subject): **same/diff** (different)/**imp** (impersonal)/ **amb** (ambiguous)
  - this tags for the same referent as the immediately previous clause
  - which means in a dialogue, the person morphology can change between speakers.
  - E.g. Maria: Qué haces? Juan: Tomo café. In this case, the psub is 'same' because the referent is Juan both times
- pp (previous pronoun): **overt/null**

Finite Verbs:

- dep(endency) type: **root** (main clause) / **sub** (dependent clause) / **rel** (relative clause)
  - -INT for questions
- subid (subject ID): the ID of the corresponding subject's token
- morphology: e.g. "me fuera": <morphology>1si-s</morphology>
  - person: 1/2/3
  - number: s/p
  - tense:
    - p=present
    - i=imperfect
    - r=preterite
    - f=future
  - aspect:
    - p=perfect
    - g=progressive
  - mood:
    - i=indicative
    - s=subjunctive
    - c=conditional
    - m=imperative